Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Application of Decision Tree (M5Tree) Algorithm for Multicrop Yield Prediction of the Semi-Arid Region of Maharashtra, India

Kalpesh S. Borse [1,*]; **Prasit Agnihotri** [1]

1. Ph. D., Department of Civil Engineering, National Institute of Technology, Surat, India

* Corresponding author: *kalpeshborse22@gmail.com*

https://doi.org/10.22115/SCCE.2024.383387.1601

ARTICLE INFO

Article history:
Received: 27 January 2023
Revised: 25 December 2023
Accepted: 24 February 2024

Keywords:
Artificial intelligence;
Climatic parameters;
M5tree;
Nasik region;
Yield prediction.

ABSTRACT

India, with a population of 1.40 billion, is the most populous country in the world, necessitating increased food production. For making decisions about issues relating to food security, reliable crop yield estimation is essential. Knowing the expected yield of one's standing crops is crucial to farmers and can be a complicated task in and of itself. Modern artificial intelligence algorithms have shown to be highly useful tools for accurately predicting agricultural production. The primary focus of this study is on predicting the yield of five important crops grown in the Nashik region. Crop yield is significantly influenced by climatic variables such as rainfall, minimum and maximum temperatures, relative humidity, and evaporation. In this regard, we used the Decision tree (M5tree) method to predict the yield of five important crops farmed in the Nashik region of Maharashtra state in India: rice, jowar, maize, groundnut, and sugarcane. With acceptable accuracy, the developed models have functioned well. The association between climatic variables and the agricultural production of the crops under study was disclosed by the decision trees, and the rule accuracy validated this relationship.

How to cite this article: Borse KS, Agnihotri P. Application of decision tree (M5Tree) algorithm for multicrop yield prediction of the semi-arid region of maharashtra, India. J Soft Comput Civ Eng 2025;9(1): 64–90. https://doi.org/10.22115/scce.2024.383387.1601

2588-2872/ © 2025 The Authors. Published by Pouyan Press.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction and related work

India is the world's most populated country, with over 1.40 billion population [1]. A total of 82% of its farmers are small and marginal, and around 70% of the country's rural population still depends on agriculture for their livelihood. Further, India is the largest consumer (27%), producer (25%), and importer (14%) of pulses. It is also the second-largest producer of fruit and vegetables in the world, contributing 10.9 % and 8.6 % respectively, of the world's output of these foods. Worldwide, India stands second in the production of rice, wheat, sugarcane, cotton, and groundnuts [2]. Feeding such a large population demands more agricultural production. However, food security is a major issue that plays an important role in this regard. The key factor in addressing the problems associated with food security is the effective management of agricultural output in order to meet basic nutritional needs. The State Governments, decision-makers, and associated organizations are of the highest significance in this respect for the establishment of new policies or specific alterations in the current policies to increase the farmer's net production and revenue. Therefore, for decision-making about issues related to food security, an accurate assessment of crop yields is crucial [3].

Further, crop yield prediction is itself a very complex process and it is very helpful and important to the farmers to know the future yield of their standing crops. The crop yield is mainly dependent on parameters such as acreage, sources of irrigation, minimum and maximum temperature, rainfall, relative humidity, evaporation, etc. The precision in crop yield prediction can be attained by implementing proper inputs and model without disturbing the agricultural production nature and systems. However, to date, very few tools or technologies are available to predict the accurate crop yield under changing climatic conditions. Many of the researchers have used on-field surveys, crop growth simulation model, data-driven models, remote sensing methods, statistical models, and their combinations for crop yield prediction. These strategies handle somewhat different elements of crop yield prediction on their own. Most of the previous research work focused on crop growth models which represent crop growth and development based on agronomic principles of plant, environment, and management interactions in field surveys [4,5] On the other hand, Remote sensing methods rely on satellite-derived imageries to determine crop stage and subsequently predict final crop yield [6]. Recent research has boosted yield forecasting models by combining several methodologies in novel ways. Lobell et al. (2015) and Zhao et al. (2020) [7,8] developed statistical models to estimate the actual yield using high-resolution remote sensing imageries and crop modeling. Likewise, Newlands et al. (2014) [9] used remote sensing, crop modeling, and statistical models to construct a probabilistic yield forecasting model for Canada. Deep learning (DL) technologies have recently been introduced for crop yield prediction. Between 2008 and 2016, Khaki and Wang [10] constructed a deep neural network model to forecast corn yield over 2,247 sites. Other approaches, such as LASSO, shallow Neural Networks, and regression trees, outperformed their model. Furthermore, Kim et al. (2019) [11] built a deep neural network model for agricultural production prediction using optimal input factors from satellite products and meteorological datasets between 2006 and 2015. Wang et al. (2018) [12] developed a deep-learning model to predict Argentina's soybean crop yields and then utilized a transfer learning technique to forecast Brazil's soybean harvests with less data. Maya Gopal PS and Bhargavi R (2019) [13] developed a hybrid MLR-ANN technique

for predicting crop yield. Another work was conducted by Paudel et al. (2021) [14] who created a machine learning baseline for big scale agricultural crop production prediction by combining the agronomic concepts of crop modeling with machine learning. They started with workflow that stressed on quality, modularity, and reusability. Three machine learning algorithms, GB, SVR, and k-Nearest Neighbors were used in their proposed workflow to predict the yield of soft wheat, spring barley, sunflower, sugar beet, and potato crops at the regional level in the Netherlands, Germany, and France. They compared the results with traditional method of yield prediction and found ML techniques accuracy much better than traditional. Shahi et al. (2021) [15] simulated Pan evaporation rate using ANN Artificial Intelligence model in Damghan Region. Khosla et al **(**2019) [16] predicted the yield of Bajra, maize, rice, and ragi using MANNs-SVR hybrid model. Kim et al (2020) [17] predicted the yields of Corn and soybean using MARS, SVM, RF, extra tree, ANN, and DNN techniques. Apart from prediction of agricultural production, Artificial Inelegance techniques are also used in solving complex hydrological problems. In this connection, Gandhi, F., & Patel, J. (2022) [18] predicted the annual and seasonal rainfall using SPI and ANFIS technique**.** Further, In order to forecast six crop yields in nine West African nations—Burkina Faso, Gambia, Ghana, Guinea, Mali, Mauritania, Niger, Senegal, and Togo—Cedric et al. (2022) [19] created three machine learning models for yams, bananas, seed cotton, cassava, rice, maize, and cassava. The forecast was created at the national level and is done throughout the year. Based on a statistical examination of the whole data, they discovered that their chosen crops are among the most consumed and farmed foods in West Africa. Kuradusenge et al (2023) [20] predicted the yield of Irish potatoes and Maize using three machine learning techniques i.e Random Forest, Polynomial Regression, and Support Vector Machine. The results suggests that Random Forest is recommended model for early crop yield prediction. Ahmed S (2023) [21] developed a software framework for estimating maize production based on a modified multi-layer perceptron. The Multi-Layer Perceptron model and the Spider Monkey Optimization method were used to forecast yield. The prediction effectiveness of the proposed MLP-SMO model was evaluated using several evaluation metrics, and the model outperformed in the prediction process with a Root-Mean-Square Error value of 0.11, which is the lowest among all techniques considered in the statistical analysis in their study. Jhajharia, K. et al (2023) [22] used various machine learning techniques (RF, SVM, GD, LSTM and LASSO regression) to estimate crop yield of seven crops grown in Rajasthan, India. The RF technique outperformed others with 0.963 R^2, 0.035 RMSE, and 0.0251 MAE.

It was found that, numerous Artificial Intelligence techniques were used for crop yield prediction modeling. It is also revealed that very few researchers have used the decision tree algorithms in the process of developing the crop yield prediction models with multiple climatic parameters. However, few studies (Khosla and Kim, Cedric, Kuradusenge, Paudel et al) have considered more than one crop in the process of developing the crop yield prediction models. Despite of considering more than one crops for yield prediction, those studies lack in the selection of multi climatic parameters for the prediction of multi crop yield. Further Jhajharia, K. et al predicted the multi crop yield but they have considered only rainfall and soil as input datasets to forecast the yield of seven crops. The fact is that the crop yield is also dependant on other climatic factors such as minimum and maximum temperature, evaporation, humidity etc. Also already developed

models accuracy for implementation is matter of concern when applicable in actual field conditions.

In this regard, an attempt has been made to develop the crop yield prediction models with considering climatic parameters for five major *Kharif* crops, namely Rice, Jowar, Maize, Groundnut, and one perennial cash crop i.e., Sugarcane grown in the Nashik region using the Decision tree (M5tree) algorithm.

2. Study area and data collection

The study mainly focuses on the Nashik region of the state of Maharashtra (India). The Nashik region lies between 19°35'6.37''N to 20°32'51''N latitude and 73°25'9.79''E to 74°22'1.15''E longitudes in the Maharashtra state of India. The total geographical area of the Nashik region is about 6376.93 KM^2. The semi-arid region of the state of Maharashtra experiences a dry season that lasts from the beginning of October until the middle of June and a rainy season that lasts from June until the beginning of October. The Nashik region receives an average of approximately 713.50 millimeters of precipitation each year, however, this number is subject to variation. The geographical location of the Nashik region is shown in Fig.1.

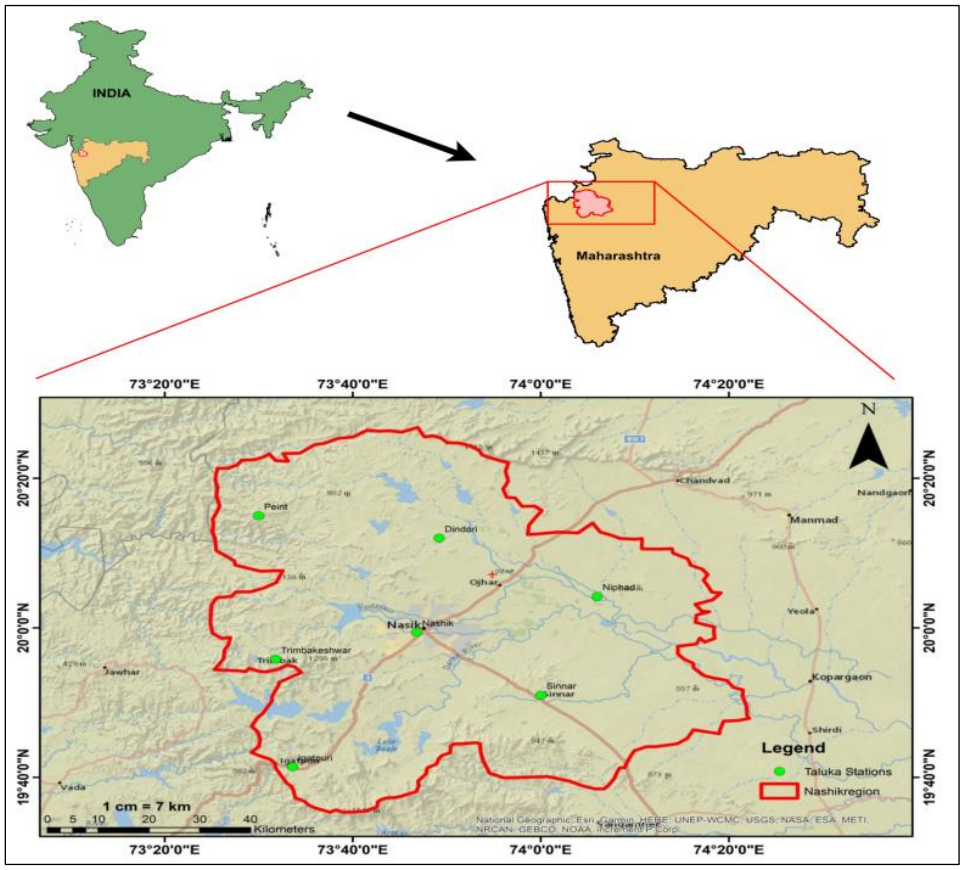


Fig. 1. Geographical location of the Nashik region.

For the present study, climatological data including daily rainfall, minimum and maximum temperature, evaporation, and relative humidity were obtained from WRDHP, Nashik

(Maharashtra) for the seven study stations *viz.* Nashik, Trimbakeshwar, Dindori, Igatpuri, Sinnar, Niphad, and Peint from 1987 to 2020, (i.e., 33 years) for the development of the individual crop yield models. Following that, crop yield statistics for all 07 stations were obtained from the Agriculture Department, Government of Maharashtra, Pune for the aforementioned period (i.e., from 1987 to 2020, 33 years). The datasets utilized for this investigation are included in table 1 below, along with their descriptions.

Table 1
Model input parameter used and their description.

No.	Parameter	Description	Units	Frequency
1	Average rainfall (R)	Average Precipitation recorded in a study area during *Kharif* season	Integer (mm)	Seasonal
2	Minimum average temperature (T_{min})	The lowermost average temperature recorded in a study area during *Kharif* season	Integer (°C)	Seasonal
3	Maximum average temperature (T_{max})	The highest average temperature recorded in a study area during *Kharif* season	Integer (°C)	Seasonal
4	Average evaporation (E)	Average moisture loss from Soil surface recorded in a study area during *Kharif* season	Integer (mm/day)	Seasonal
5	Average relative humidity (Rh)	The ratio of the current average absolute humidity to the highest average possible absolute humidity	%	Seasonal
6	Average crop yield (Y)	The production of grains, oil seed, and cash crops after harvesting	Kg/ha	Seasonal

3. Methodology

The major goal of this work is to use the M5tree algorithm to construct agricultural yield prediction models for each *Kharif* crop, including rice, jowar, maize, groundnut, and one cash crop, sugarcane, grown in the study area. The obtained climatological and crop yield data were statistically verified for outliers prior to analysis. At the level of 95% significance, outliers were examined. The current outliers were then eliminated.

3.1. Input parameter selection

Although several factors affect the crop yield, we tried to develop individual crop yield models by considering the maximum effect of climatic parameters such as Average Rainfall (R), Minimum Average Temperature (T_{min}), Maximum Average Temperature (T_{max}), Average Evaporation (E), and Relative Humidity (Rh) on the crop growth and its yield. However, for developing yield prediction models using novel AI techniques often require large amounts of training data to perform well, Durrant et al., (2021) [23]. In this regard, we tried to increase the no. datasets of each input parameter, as well as observed yield by using normal and inverse normal distributions as the availability of the seasonal datasets are not enough to achieve the desired model accuracy. To increase the amount of data in AI modeling using normal and inverse normal distributions, the following popular techniques can be used.

- Sampling from the distribution: One simple way to increase the amount of data is to generate more samples from the distribution. For example, if you have a normal

distribution with mean μ and standard deviation σ, you can generate more samples by using a random number generator to draw values from the distribution.

- Data augmentation: Another approach is to use data augmentation techniques to create more data from the existing dataset. For example, you can use transformations such as rotation, scaling, or cropping to create new images from existing ones.
- Interpolation: You can also use interpolation techniques to generate more data. For example, you can interpolate between existing data points to create new ones. This approach can be particularly useful when working with time series data.
- Monte Carlo simulations: Monte Carlo simulations involve generating random numbers to simulate complex systems. This approach can be used to generate more data for AI modeling. For example, you can use Monte Carlo simulations to simulate the behavior of a complex system, such as the stock market, and generate more data for training and testing.

In this study, we have used the distribution method for increasing data using normal and inverse normal distributions with equations.

Suppose we have a normal distribution with mean μ and standard deviation σ. The probability density function (PDF) of the normal distribution is given by:

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) * e^{\frac{-(x-\mu)^2}{(2\sigma^2)}} \quad (1)$$

where x is a random variable, μ is the mean of the distribution, σ is the standard deviation of the distribution, e is the base of the natural logarithm, and π is the mathematical constant pi.

Similarly, we can generate samples from an inverse normal distribution using the inverse Cumulative Distribution Function (CDF) method. The inverse normal distribution, also known as the Gaussian inverse, is the inverse of the standard normal distribution. The PDF of the inverse normal distribution is given by:

$$f(x) = \left(\frac{1}{\sqrt{2\pi x^3}}\right) * e^{-\left(\frac{1}{2x}\right)*\left(\frac{(ln(x)-\mu)^2}{\sigma^2}\right)} \quad (2)$$

where x is a random variable, μ is the mean of the distribution, σ is the standard deviation of the distribution, e is the base of the natural logarithm, and π is the mathematical constant pi.

To generate samples from this distribution, we used the inverse CDF method. The CDF of the inverse normal distribution is given by:

$$\mathrm{F(x)} = \int [0, \mathrm{x}]\mathrm{f(t)dt} \quad (3)$$

Where, f(t) is the PDF of the inverse normal distribution. We can use numerical methods to calculate the CDF, or we can use lookup tables that provide pre-calculated values of the CDF for a range of values of x. Once we have the CDF, we can use the inverse CDF method to generate samples from the distribution. The inverse CDF method involves generating a random value u between 0 and 1 and then finding the value x such that F(x) = u. The value x is the generated sample. After performing the normal distributions, the no. of datasets generated for each crop is

represented in the below table. To achieve the maximum accuracy in developing the AI based crop yield prediction models we have used the seasonal average of all the input parameters.

Table 2
No. of datasets generated for each input parameter of each crop.

No	Crop	No. of datasets
1	Kharif Rice	213
2	Kharif Jowar	66
3	Kharif Maize	52
4	Kharif Groundnut	215
5	Sugarcane	155

The time series plots of seasonal average datasets generated after performing the normal distributions is presented in the following figures from 2 to 7 for each crop viz Kharif rice, jowar, maize, groundnut, and sugarcane respectively.

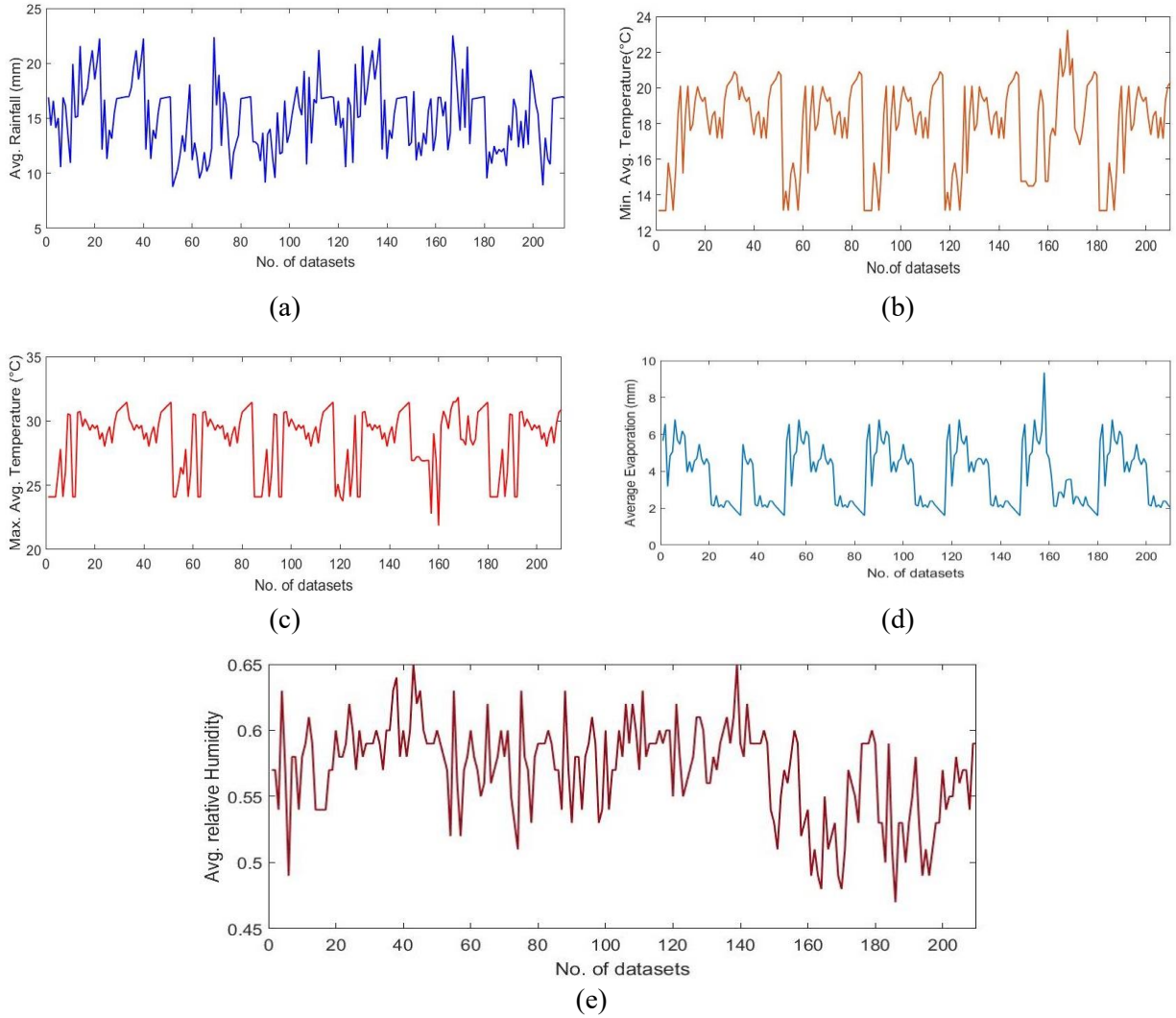


Fig. 2. Time series of seasonal datasets generated after performing the normal distributions in respect of Kharif rice. (a) avg. rainfall; (b) min. avg. temperature c. max. avg. temperature d. avg. evaporation and e. avg. relative humidity)

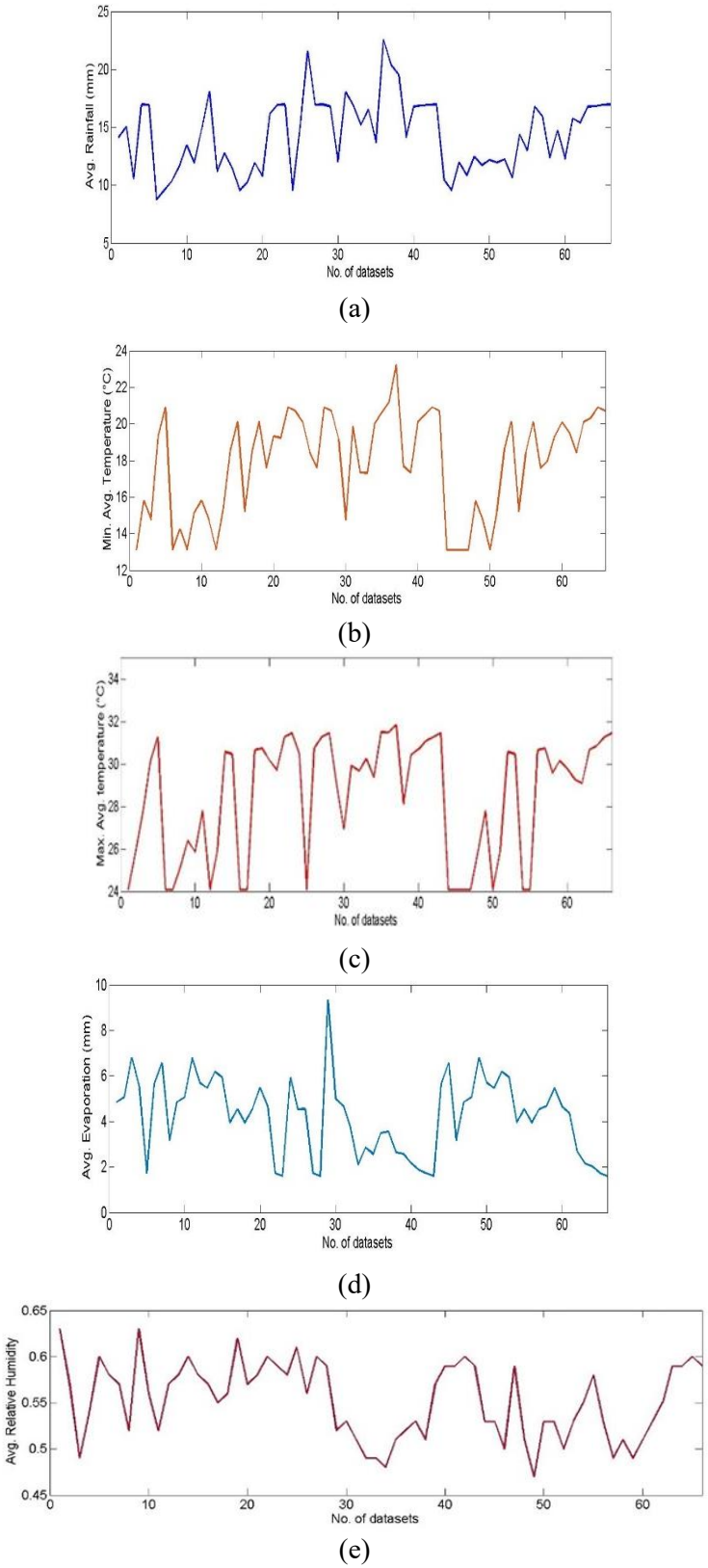


Fig. 3. Time series of seasonal datasets generated after performing the normal distributions in respect of Kharif jowar. (a) avg. rainfall (b) min. avg. temperature (c) max. avg. temperature (d) avg. evaporation and (e) avg. relative humidity.

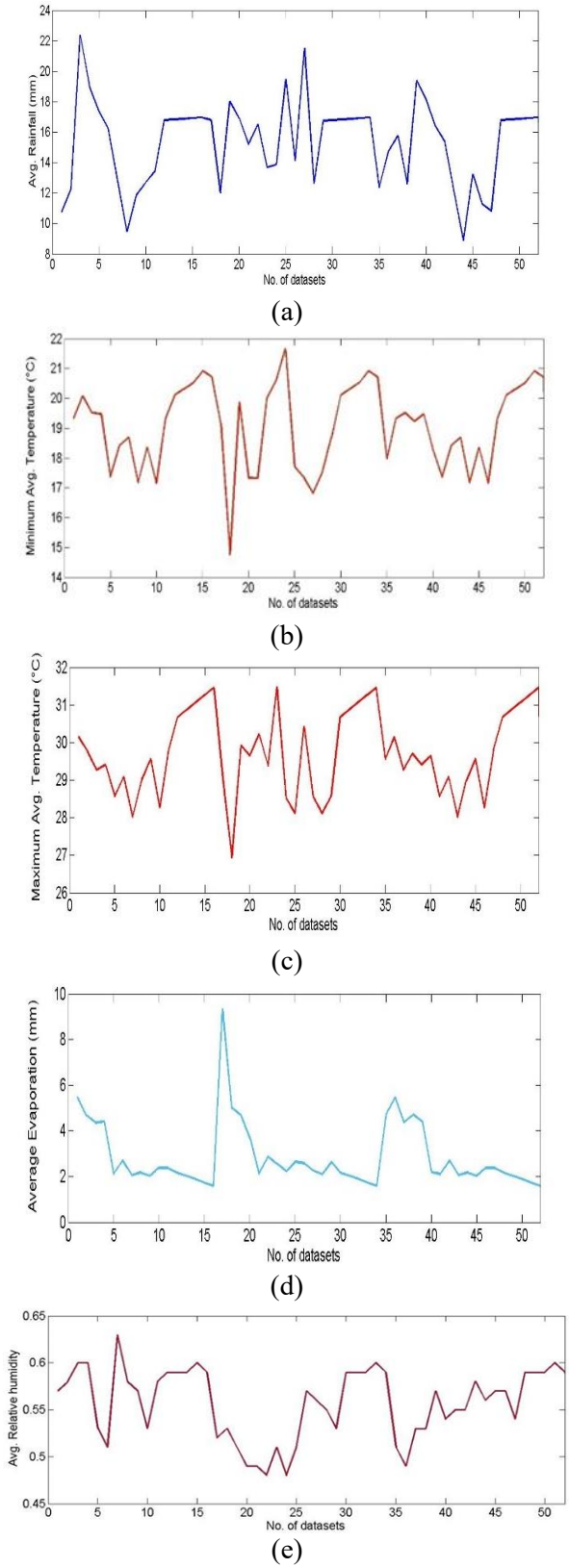


Fig. 4. Time series of seasonal datasets generated after performing the normal distributions in respect of Kharif maize. (a) avg. rainfall (b) min. avg. temperature (c) max. avg. temperature (d) avg. evaporation and (e) avg. relative humidity.

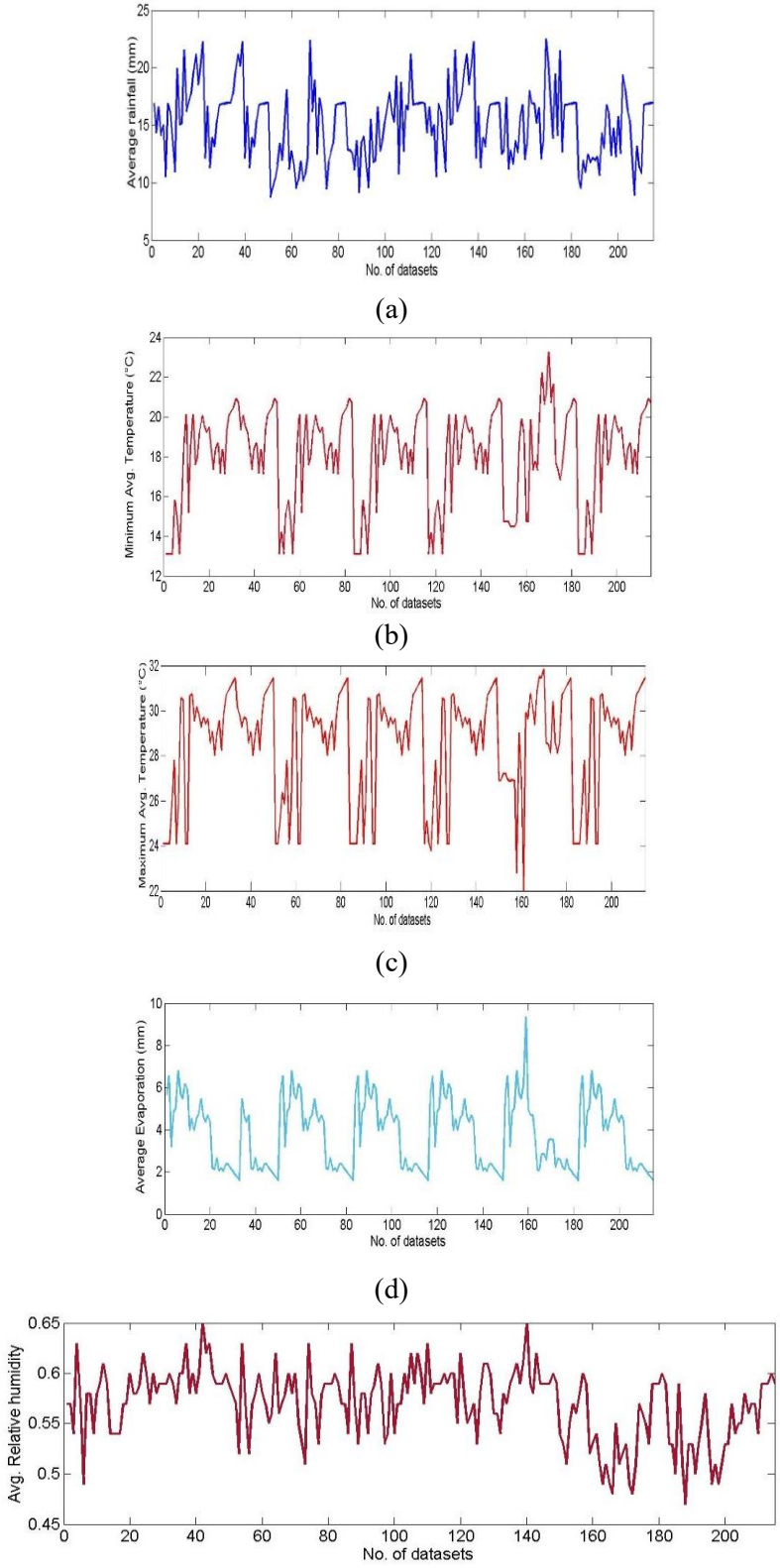


Fig. 5. Time series of seasonal datasets generated after performing the normal distributions in respect of Kharif groundnut. (a) avg. rainfall b. min. avg. temperature (c) max. avg. temperature (d) avg. evaporation and (e) avg. relative humidity.

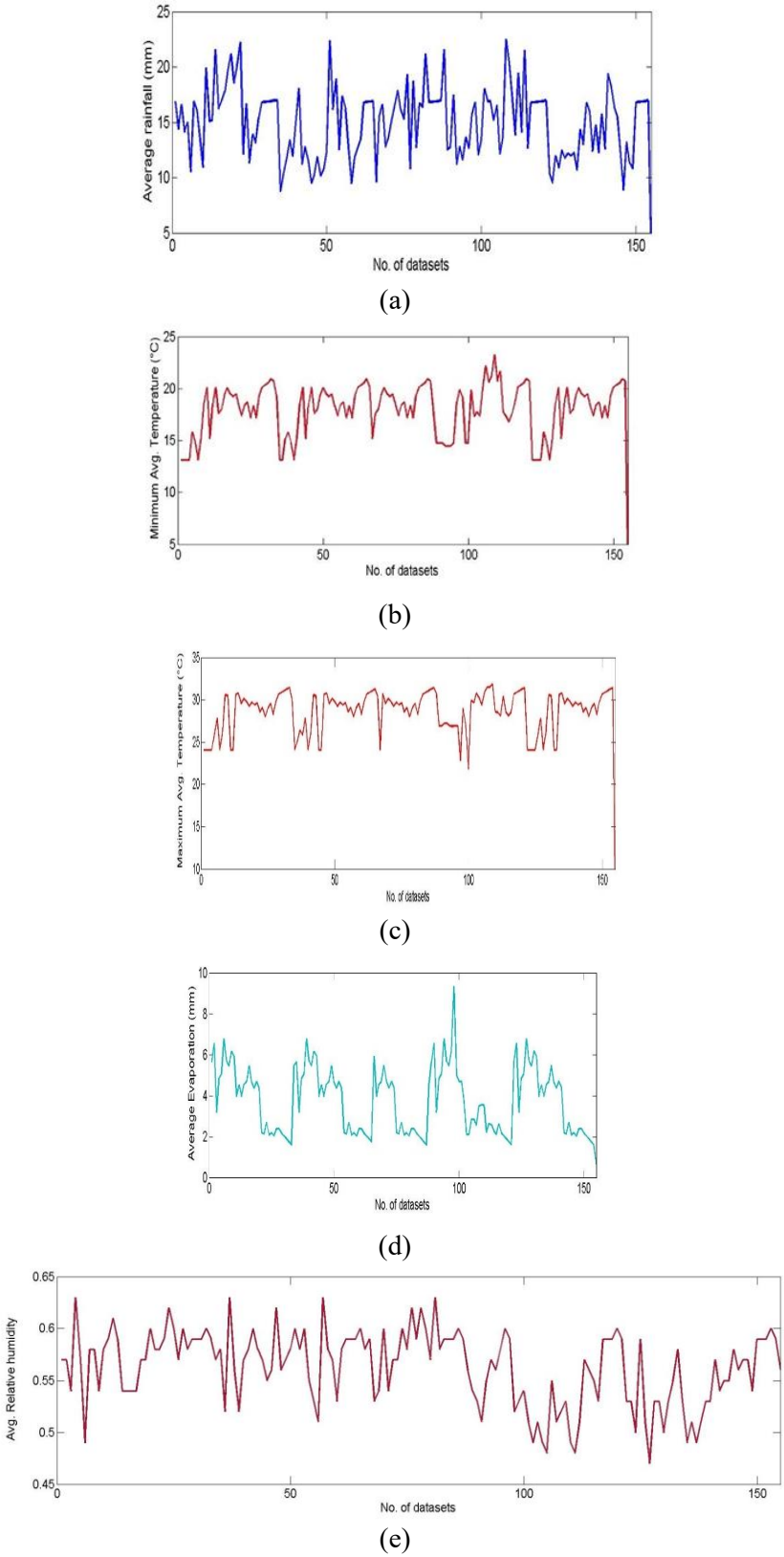
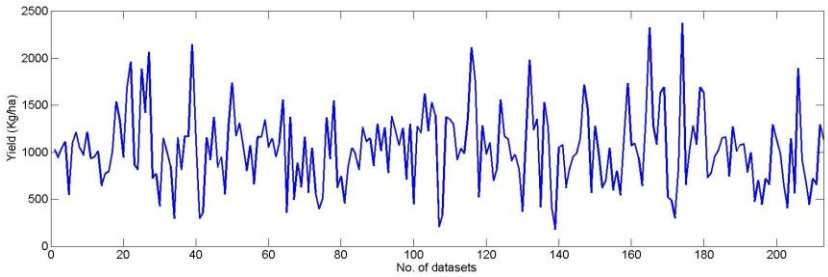
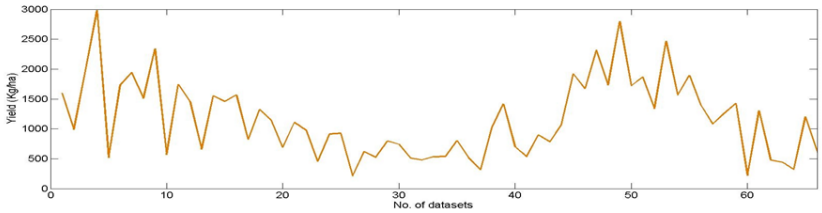


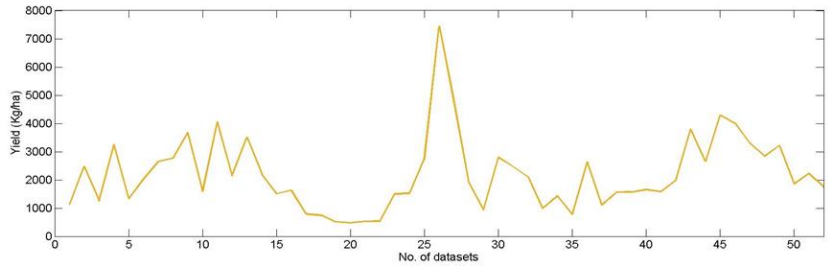
Fig. 6. Time series of seasonal datasets generated after performing the normal distributions in respect of sugarcane. (a) avg. rainfall (b) min. avg. temperature (c) max. avg. temperature (d) avg. evaporation and (e) avg. relative humidity.



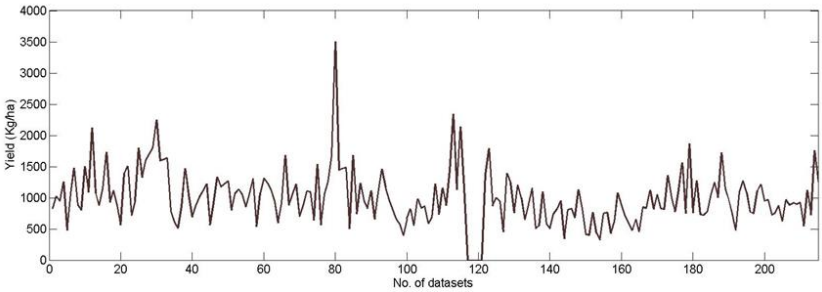
(a) Kharif rice.



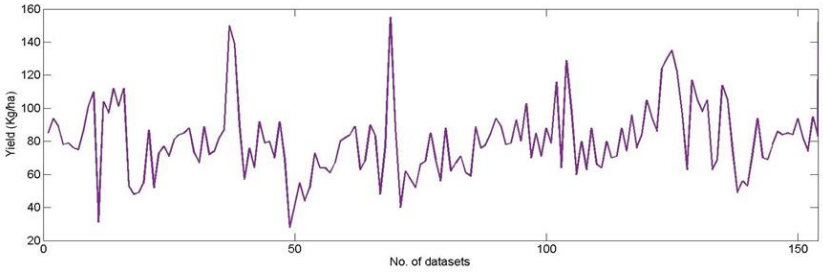
(b) Kharif jowar.



(c) Kharif maize.



(d) Kharif groundnut.



(e) Sugarcane.

Fig. 7. Time series of actual seasonal crop yield datasets generated after performing the normal distributions. (a) Kharif rice (b) Kharif jowar (c) Kharif maize (d) Kharif groundnut and e. Sugarcane.

Moreover, similar technique (normal distribution) has been adopted to generate the actual yield datasets for each crop. The time series plots of actual seasonal average crop yield datasets generated after performing the normal distributions is presented in the figure 7 for each crop viz Kharif rice, jowar, maize, groundnut, and sugarcane respectively.

A proposed methodology for the development of crop yield models for the five major crops using the M5 tree algorithm is shown in Fig. 8.

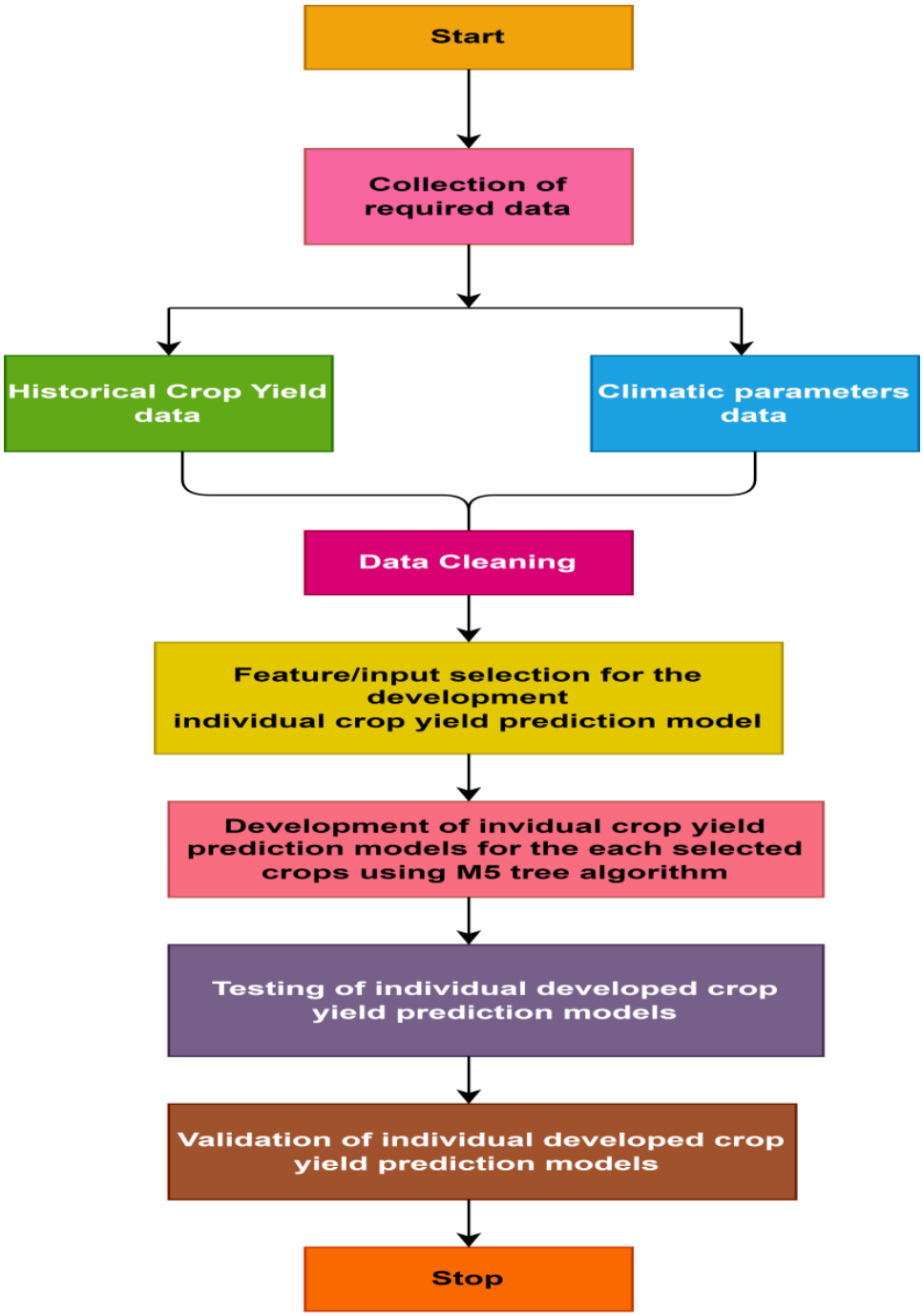

Start
Collection of required data
Historical Crop Yield data
Climatic parameters data
Data Cleaning
Feature/input selection for the development individual crop yield prediction model
Development of invidual crop yield prediction models for the each selected crops using M5 tree algorithm
Testing of individual developed crop yield prediction models
Validation of individual developed crop yield prediction models
Stop

Fig. 8. Flowchart of proposed methodology for the development of crop yield models.

3.2. Decision tree (M5tree)

The M5 tree imparts a tree categorization approach to quantify the relationship between independent and dependent variables, [24]. It is corresponding to separated linear functions, which are a combination of linear regression and tree regression models and are widely used in many fields. The tree regression model separates the data range into sub-areas known as leaves, and each leaf is given a number. The regression model produces a regression equation for the full data sets. According to Nourani et al. (2019) [25] the M5tree predicts or computes continuous numerical variables by replacing the linear regression equation with a label in nodes. Figure 9 depicts the overall structure of the M5tree.

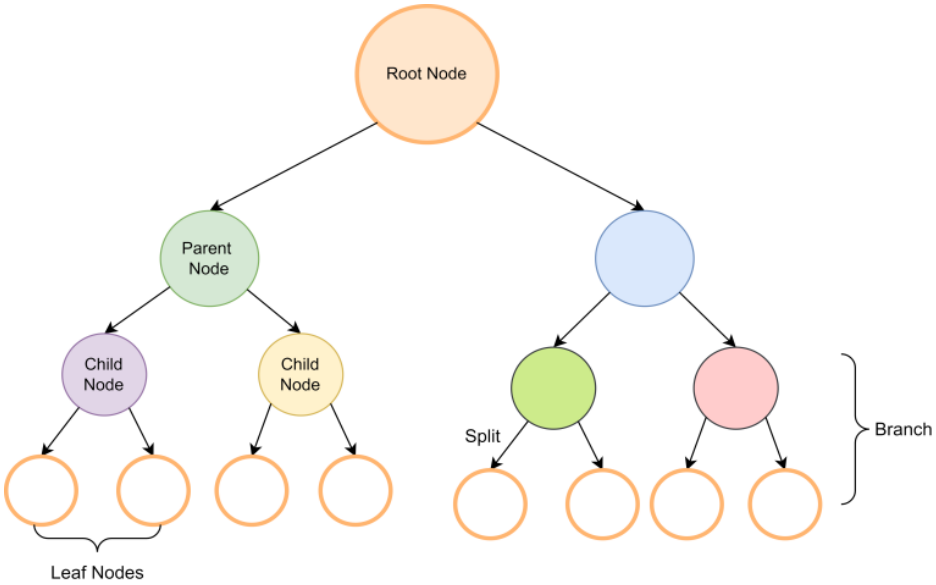


Fig. 9. General structure of the Decision tree (M5tree).

3.3. Software use of decision tree (M5tree)

The fundamental goal of this technique is to use decision rules drawn from training data to develop a model that predicts a class or a target variable. To tackle this problem, a tree-like structure with nodes specified from the root to the leaf is used. A classification or regression tree is constructed based on the categories of variables (discrete or continuous).

It is a regression decision tree learner used to predict numerical response variable values. It is a binary decision tree with linear regression functions at the leaf nodes that may predict continuous numerical features. It is capable of recreating events with hundreds of attributes and exceptionally high complexity. This feature distinguishes M5tree from other regression tree learners at the time (such as MARS), whose processing cost increases rapidly as the number of features increases. There are two steps in developing model tree. Making a decision tree with a splitting rule is the initial stage. The splitting criterion of the M5tree technique is based on the standard deviation of the class values that arrive at a node.

When data is partitioned, the standard deviation values in child nodes are lower than those in the parent node. After assessing all potential splits, M5tree selects the split that improves the

expected error reduction. This split typically produces a massive tree-like structure, which can lead to overfitting. As a result, the tree has to be cut down. and, the enormous tree would be reduced in the second step, and the subtrees would be replaced by linear regression functions. The model tree is constructed by subdividing the parameter space and running a linear regression model in each of them.

3.4. Procedure of development of multi-crop yield prediction models using decision tree (M5Tree) algorithm

A regression-based decision tree was built in this study to predict agricultural production based on climatological variables. The Classification and Regression Tree model was built expressly for this purpose. The data is divided several times by the decision tree until the stopping criterion is meet. Non-overlapping sections were created iteratively in the training dataset, and each of these regions has a decision rule connected to it that is in charge of categorizing data in that region. A correct decision tree may be viewed as a collection of intelligible decision rules. Each path from a root to a leaf node may be transformed into an "IF-THEN" sequence that can be coupled to observations to categorize the objects. Figure 10 shows a flowchart depicting the step-by-step functioning of the decision tree (M5 tree) method.

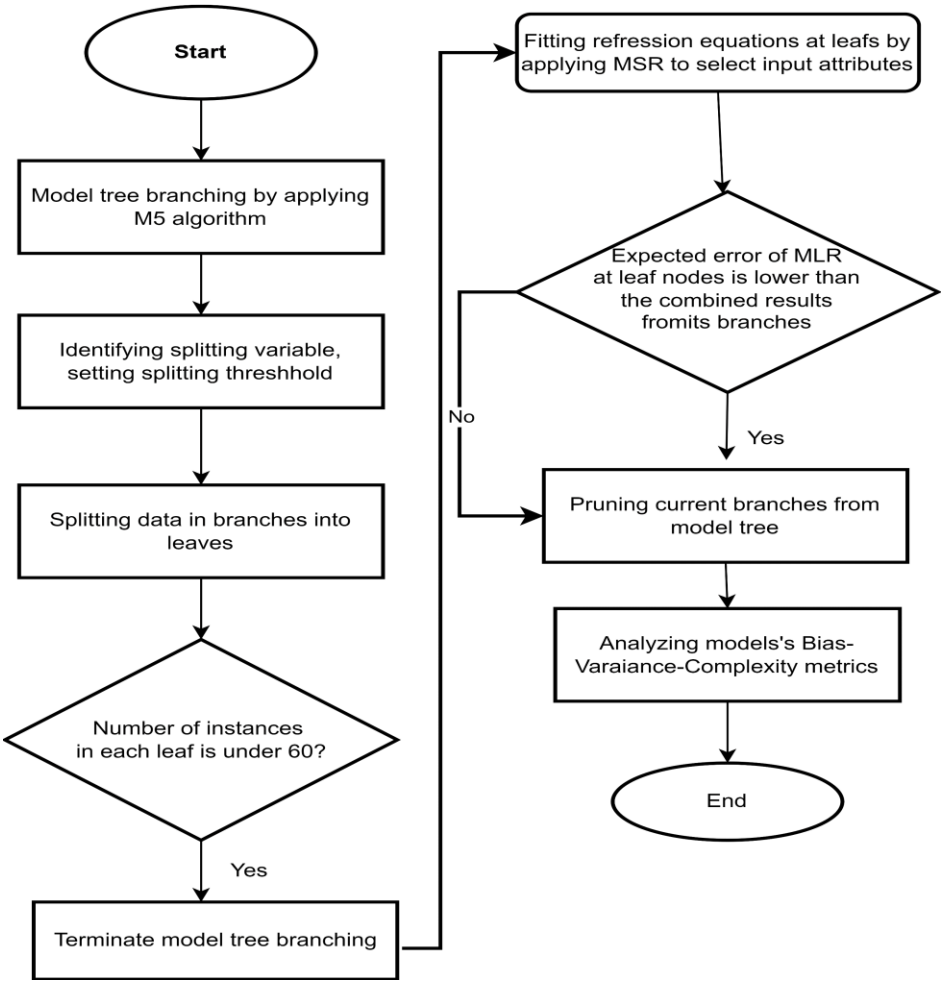

Start
Model tree branching by applying M5 algorithm
Identifying splitting variable, setting splitting threshhold
Splitting data in branches into leaves
Number of instances in each leaf is under 60?
Yes
Terminate model tree branching
Fitting refression equations at leafs by applying MSR to select input attributes
Expected error of MLR at leaf nodes is lower than the combined results fromits branches
No
Yes
Pruning current branches from model tree
Analyzing models's Bias-Varaiance-Complexity metrics
End

Fig. 10. Step-by-step working of the decision tree (M5Tree) algorithm.

The decision tree approach consists of three steps: training, splitting, and stopping criteria.

Step 1: Training

Two key steps were explored throughout. There must be a growth period, and trimming is optional. During the growth stage, the tree is built from the top down, starting with the root. The dataset is regularly separated based on the attributes and partitioning criteria.

Step 2: Splitting

This is optional, and its main purpose is to reduce over-fitting. The process is continued until the best sub-tree is identified. The growth phase should be limited early at a few nodes to minimize the explosion of tree complexity. Several trials have been carried out in order to establish the best sub-tree for each model. In this case, we did 100 to 110 trials for each model to discover the best sub-tree.

Step 3: Stopping criteria

Limiting the number of samples per node or utilizing the least amount of impurity to divide the node further is the most typical stopping condition. In this instance, we reduced the node splitting further till the required correlation was attained. A thorough MATLAB code has been constructed to develop the yield prediction models for each crop using the decision tree (M5tree) technique.

4. Results and discussion

For yield prediction of various crops, the developed decision tree (M5tree) models have selected the input parameters based on co-variance. It means that the factors which affect crop yield adversely would be selected while training the models. All the developed models with their performance are discussed in the subsequent sections. For the development of each model, 80 % of the data was used for calibration whereas 20 % of the total dataset was used for validation.

4.1. Prediction of the Kharif rice yield

To predict the yield of Rice in the Nashik region, high-impacting parameters were selected by the model covariantly. From fig. 11, it can be seen that average rainfall and evaporation are the most important features in determining the yield of rice.

For the development of the rice yield prediction model, 173 data records have been used for the calibration of the model while 40 data records were used for the validation purpose. Fig. 12 depicts the developed model performance during calibration and validation in respect of rice. Some of the data points not following trend line during calibration, this is because no. of trees formed were more than usual during training despite of pruning. This over fitting problem has been addressed during validation.

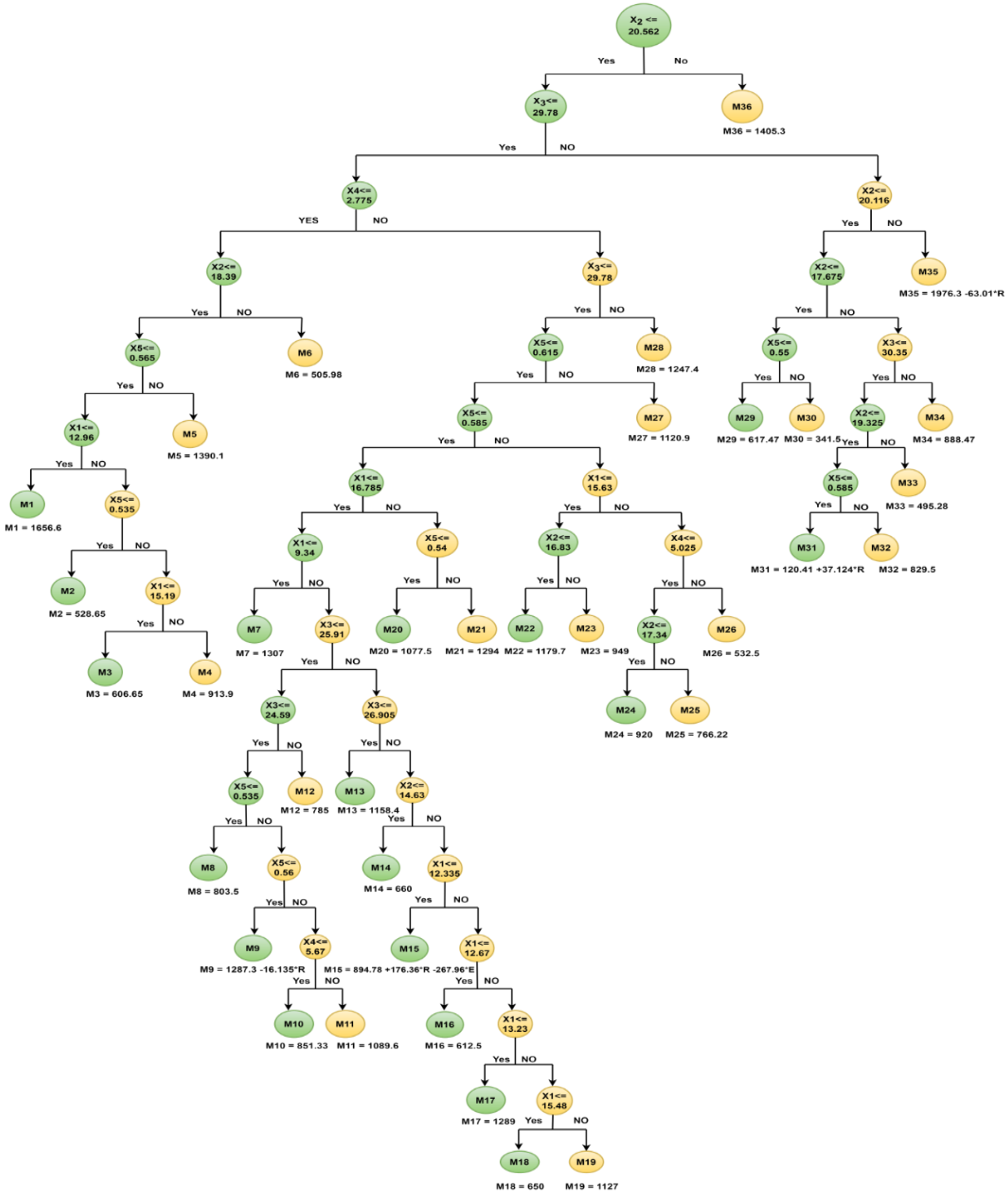
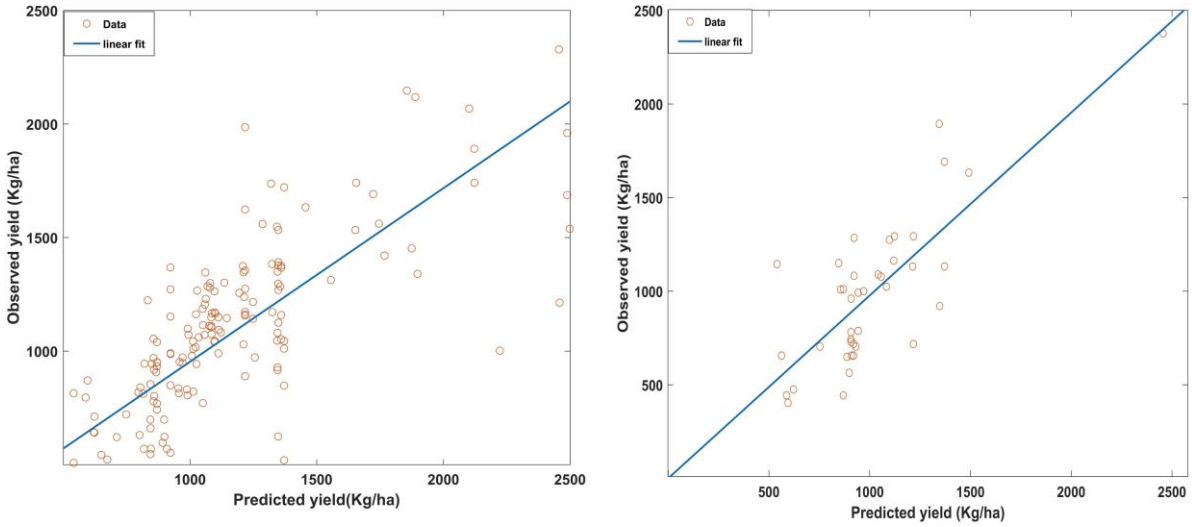


Fig. 11. M5 tree model for *Kharif* Rice.

4.2. Prediction of the Kharif jowar yield

To predict the yield of Jowar in the Nashik region, a similar methodology has been adopted. High-impacting climatic parameters on yield were selected by the model covariantly. From fig.

13 it can be seen that average minimum temperature and rainfall are most influencing parameters in determining the crop yield of Jowar.



(a) Model performance during calibration. (b) Model performance during validation.

Fig. 12. Prediction of the *Kharif* Rice yield.

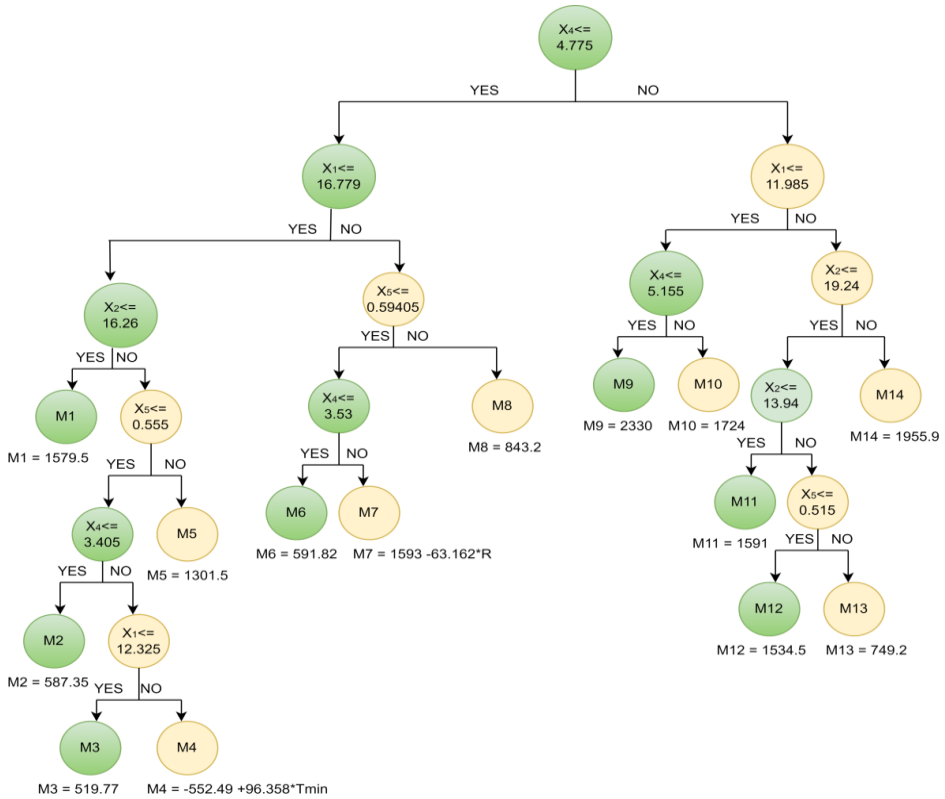
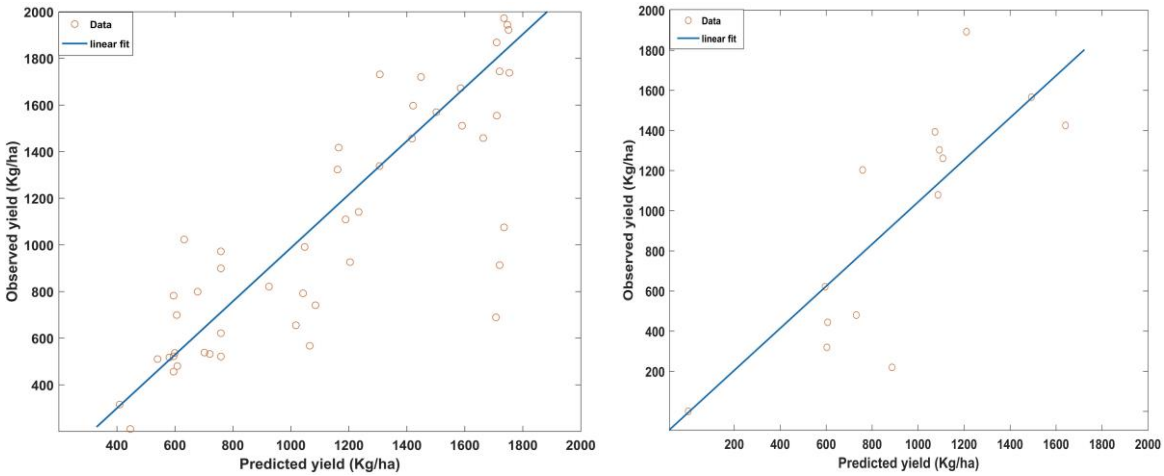


Fig. 13. M5 tree model for Kharif jowar.

For the development of the Jowar yield prediction model, 53 data records have been used for the calibration of the model while 14 data records were used for the validation purpose. Fig. 14 depicts the developed model performance during calibration and validation in respect of Jowar.



(a) Model performance during calibration. (b) Model performance during validation.

Fig. 14. Prediction of *Kharif* jowar yield.

4.3. Prediction of the Kharif maize yield

To predict the yield of Maize in the Nashik region, a similar methodology has been adopted. High-impacting climatic parameters on yield were selected by the model covariantly. From fig. 15 it can be seen that average rainfall, relative humidity, and evaporation are most influncing parameters in determining the crop yield of of maize.

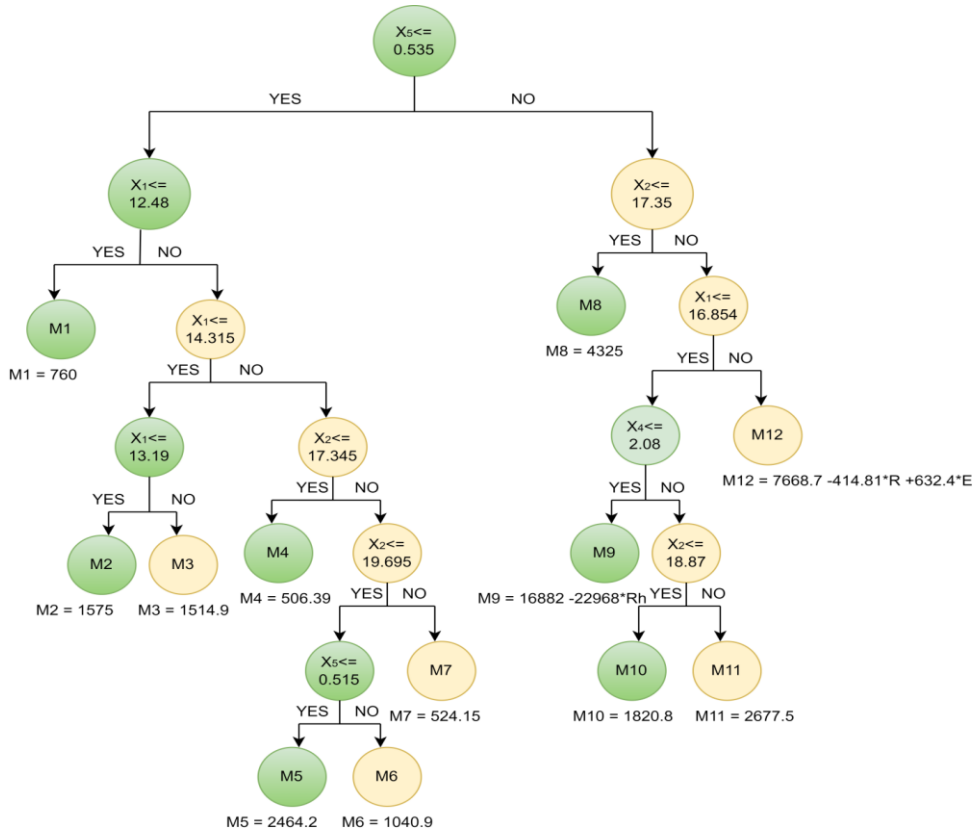
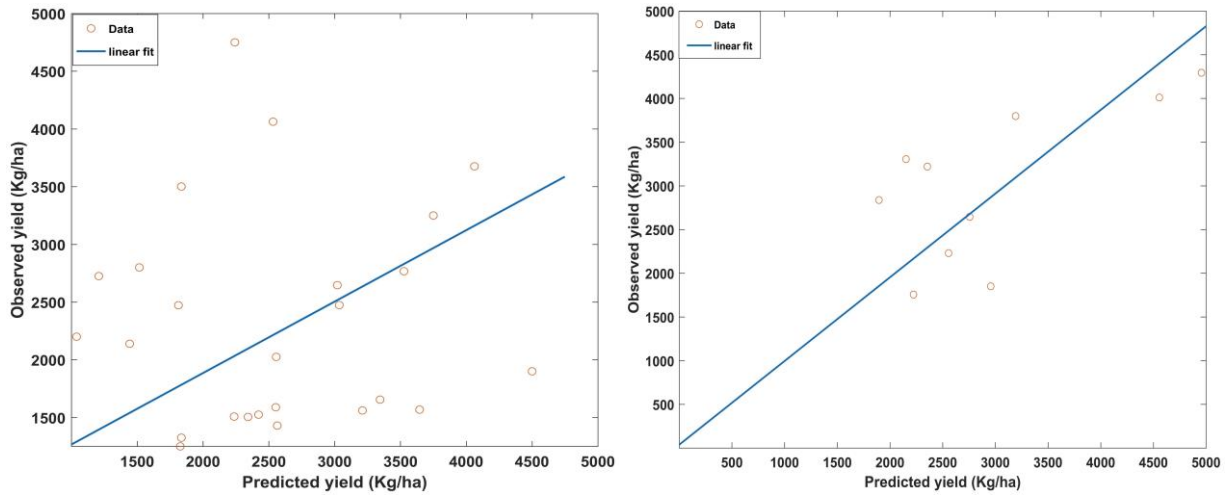


Fig. 15. M5 tree model for *Kharif* maize.

For the development of the maize yield prediction model, 42 data records have been used for the calibration of the model while 10 data records were used for validation purposes. Fig. 16 depicts the developed model performance during calibration and validation in respect of jowar.



(a) Model performance during calibration. (b) Model performance during validation.

Fig. 16. Prediction of *Kharif* maize yield.

4.4. Prediction of the Kharif groundnut yield

To predict the yield of Groundnut in the Nashik region, a similar methodology has been adopted. High-impacting climatic parameters on yield were selected by the model covariantly. From fig. 17 it can be seen that average rainfall and minimum temperature are most influncing parameters in determining the crop yield of groundnut.

For the development of the groundnut yield prediction model, 185 data records have been used for the calibration of the model while 46 data records were used for validation purposes. Fig. 18 depicts the developed model performance during calibration and validation in respect of groundnut.

4.5. Prediction of the sugarcane yield

To predict the yield of Sugarcane in the Nashik region, a similar approach has been adopted. High-impacting climatic parameters on yield were selected by the model covariantly. From fig. 19 it can be seen that average rainfall, relative humidity, and maximum temperature are most influencing parameters in determining the crop yield of of sugarcane.

For the development of the sugarcane yield prediction model, 124 data records have been used for the calibration of the model while 31 data records were used for validation purposes. Fig. 20 depicts the developed model performance during calibration and validation in respect of sugarcane.

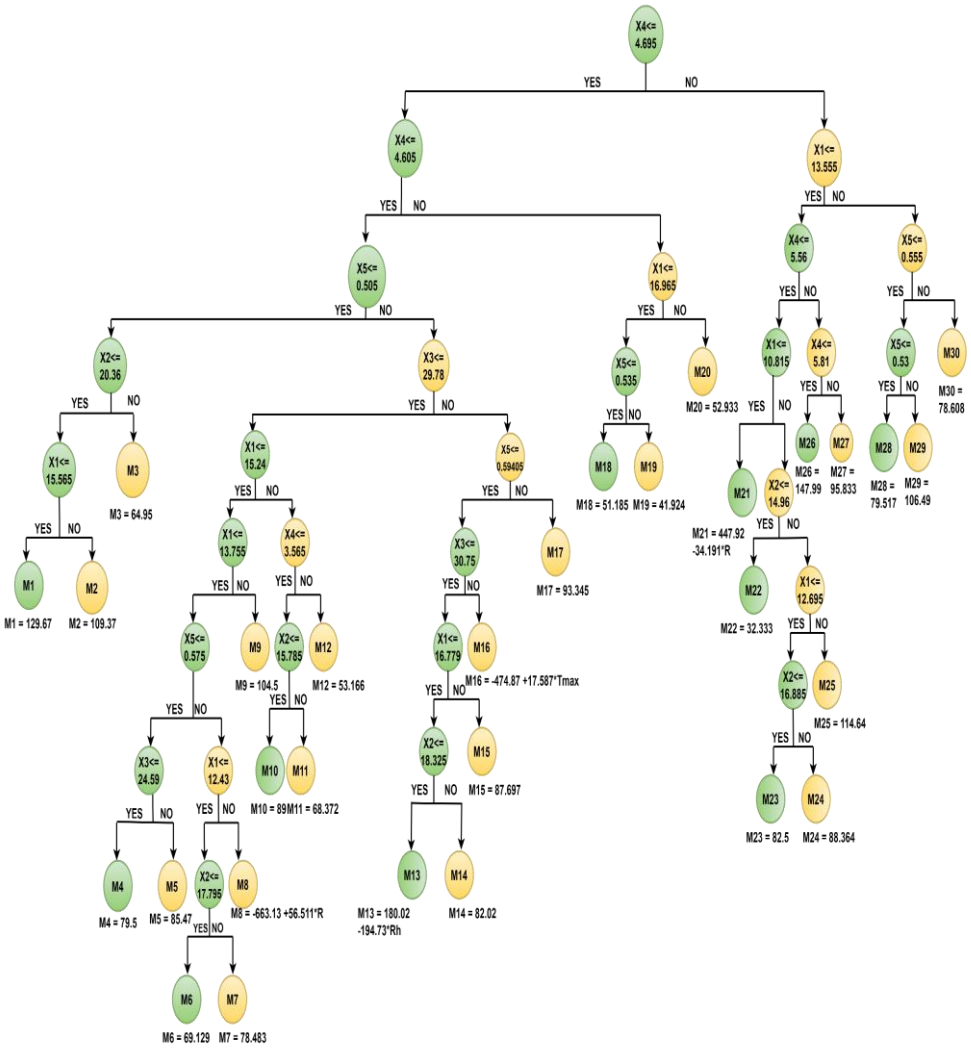
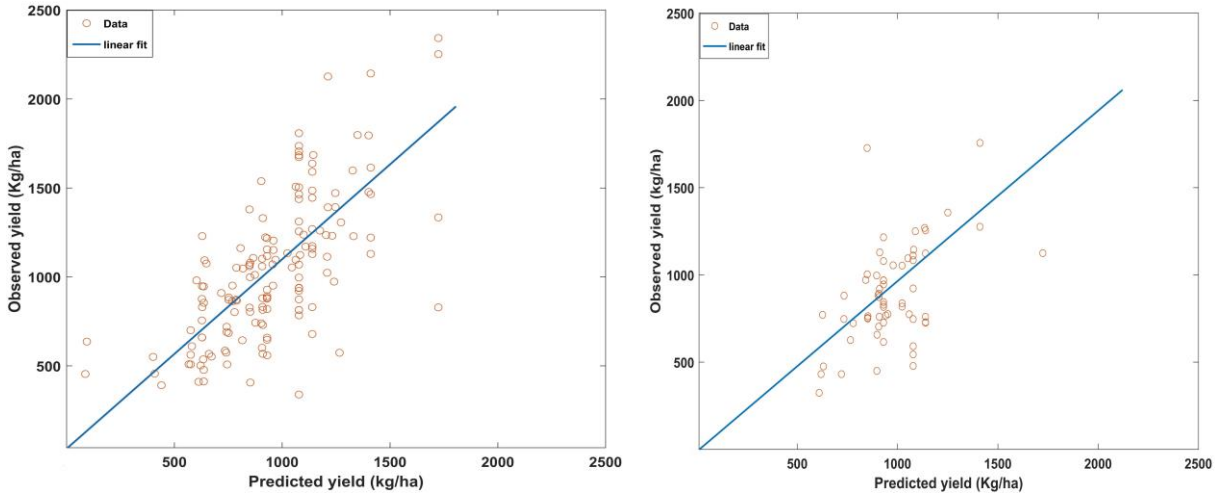


Fig. 17. A M5 tree model for Kharif groundnut.



(a) Model performance during calibration. (b) Model performance during validation.

Fig. 18. Prediction of *Kharif* groundnut yield.

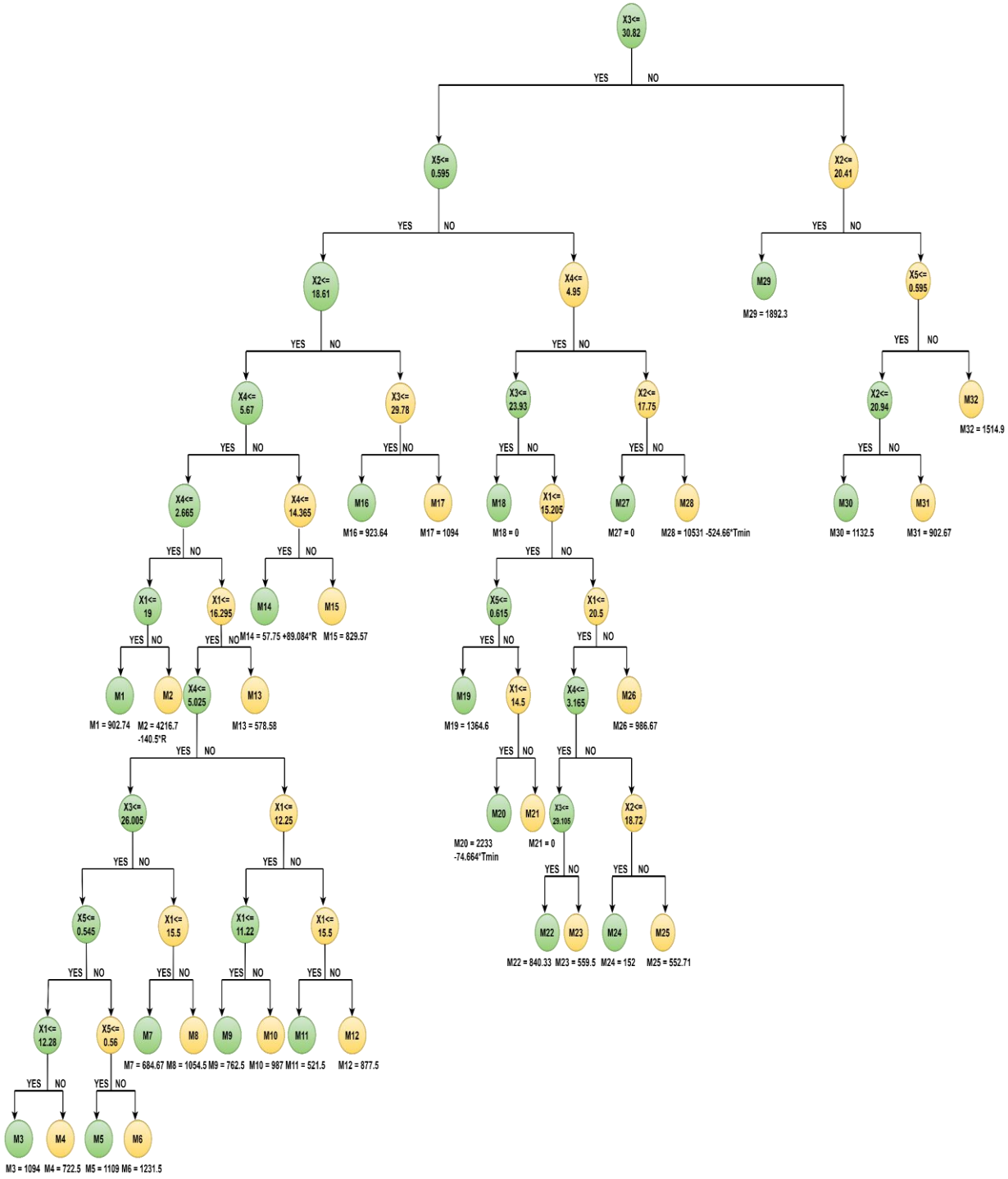
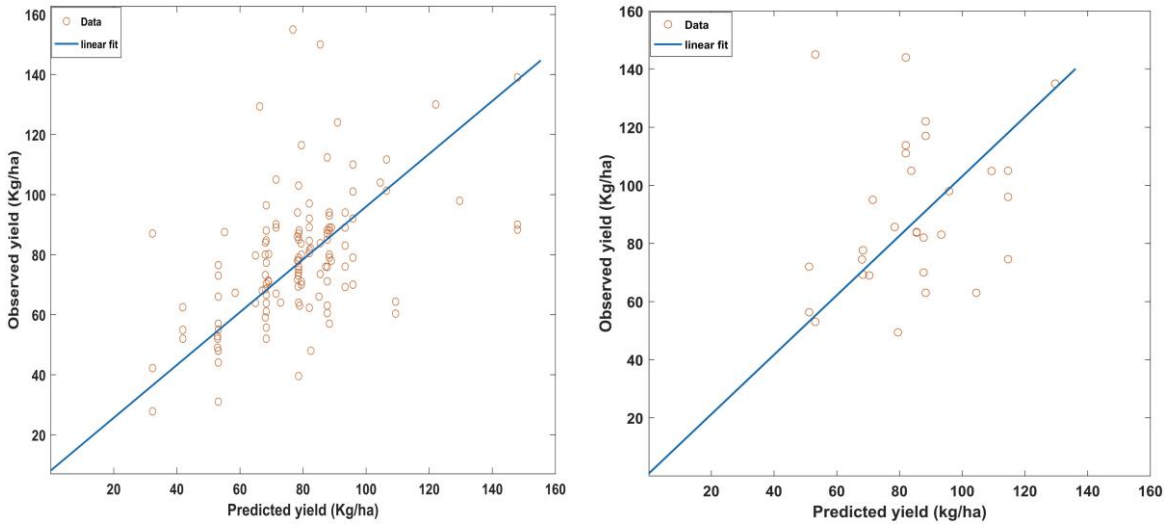


Fig. 19. A M5 tree model for sugarcane.

From the analysis, it was found that more than two variables were required to predict the individual crop yield with 36, 14, 12, 32, and 30 trees after pruning final trees for *Kharif* Rice, Jowar, Maize, Groundnut, Sugarcane respectively.



(a) Model performance during calibration. (b) Model performance during validation.

Fig. 20. Prediction of sugarcane yield.

4.6. Statistical measures for obtaining results and performance evaluation of developed models

The model's performance is evaluated using the metrics and dataset. The performance efficacy of the developed M5 tree algorithm-based models for estimating crop yields of five main crops under diverse climatic conditions is tested.

The following performance measures were used to assess the developed models:

Root Mean Square Error (RMSE)

The model's performance is evaluated using the metrics and dataset. The standard deviation of the residuals is referred to as the RMSE (prediction errors). The RMSE displays how distant the data points are from the regression line by measuring how evenly the residuals should be distributed. To put it another way, it represents how densely the data is clustered around the line of best fit. The ideal model has an RMSE of zero.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

Where, n is the sample size, $\hat{y}_i$ is the actual target value, y_i is the predicted target value.

Nash Sutcliffe Efficiency (NSE)

It evaluates the performance of a model. It might vary from -infinity to one. The efficiency of one (E = 1) shows that the predicted discharge fits the observed data precisely. When the observed mean outperforms the model, or when the residual variance (given by the numerator in the formula above) is larger than the data variance, the efficiency is zero (E = 0). (described by the denominator). The more precise the model, the closer the model efficiency is to one.

$$NSE = 1 - \frac{(Qobs - Qsim)^2}{(Qobs - Qmean)^2} \tag{5}$$

Where, $Qobs$ = Observed value of crop yield, $Qsim$= simulated value of crop yield and $Qmean$ = mean value of crop yield

Correlation (R)

Correlation coefficients are used to measure how strong a relationship is between two variables. A correlation equal to one is the best model.

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (6)$$

Where x is the observed value of crop yield, and y is the simulated value of crop yield

Coefficient of Determination $(R)^2$

It is used to measure how the developed model have strong relationship between observed and predicted. A correlation between variables equals to one is the best model.

$$R^2 = \left[\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}\right]^2$$

Where x is the observed value of crop yield, and y is the simulated value of crop yield

Table 3 depicts the performance metrics of each developed model during the calibration. From the table, among all the developed models, the sugarcane model has the highest accuracy in terms of correlation achieved between observed and predicted yield. However, the rice model accuracy is lowest during the calibration of each model.

Table 3
Performance matrix of the developed models during calibration.

No.	Crop	RMSE	NSE	Correlation	Coefficient determination
1	Kh. Rice	0.273	0.82	0.75	0.77
2	Kh. Jowar	0.374	0.71	0.82	0.83
3	Kh. Maize	0.264	0.60	0.77	0.80
4	Kh. Groundnut	0.330	0.81	0.75	0.78
5	Sugarcane	0.009	0.75	0.90	0.93

Table 4 depicts the performance metrics of each developed model during the validation. From the table, among all the developed models, the sugarcane model accuracy is the highest whereas the rice model accuracy is lowest during the validation of each model.

Table 4
Performance matrix of the developed models during validation.

No.	Crop	RMSE	NSE	Correlation	Coefficient Determination
1	Kh. Rice	0.163	0.76	0.86	0.89
2	Kh. Jowar	0.242	0.67	0.74	0.77
3	Kh. Maize	0.189	0.71	0.77	0.80
4	Kh. Groundnut	0.174	0.90	0.81	0.83
5	Sugarcane	0.013	0.88	0.89	0.91

5. Summary and conclusions

The current study is mainly focused on the application of the decision tree (M5tree) to develop the yield prediction models for five major and important crops cultivated in the Nashik region. The study illustrated the potential application of this method for deriving pertinent knowledge from secondary data on climatic parameters which are already in existence.

The M5tree demonstrates the relationship between the crop yield of the crops under examination and the climatic factors, as well as the influence of these variables on the crop yield. The decision tree analysis indicated that the yield prediction of *Kharif* rice, jowar, maize, groundnut, and sugarcane was mostly influenced by the average rainfall followed by average temperature, relative humidity, and evaporation.

Rainfall is a common factor influencing the yield of all the crops under consideration. This is because the *Kharif* crop's growth and yield are highly dependent upon rainfall. Also, the water requirement of sugarcane is the highest. The decision tree has proven a very effective technique to predict crop yield by only considering the effect of climatic parameters on it.

In compare to the models developed for multi crop yield prediction in earlier studies our models have only consider the climatic variables to predict the yields of five important crops. But a lot of other things, including the amount of fertilizer used, irrigation, the amount of water needed, etc., also affect the yield of some crops. Prediction accuracy can be improved if those factors are considered. Even if there is room to increase prediction accuracy, our models produced satisfactory results when the climate factors are merely considered.

However, because the models were developed using climatic parameters, only values of climatic parameters, such as rainfall, minimum and maximum temperatures, relative humidity, and evaporation, need to be considered when determining whether the model would produce the desired results.

Acknowledgments

The authors wish to thank the State Water Data Center (SWDC), Hydrology Project, Government of Maharashtra, Nashik, and the Chief Statistical Officer, Department of Agriculture, Government of Maharashtra, Pune for supplying meteorological and agricultural yield data respectively. The authors would also like to thank the Department of Civil Engineering, National Institute of Technology, Surat for providing the necessary facilities and infrastructure to carry out this work.

Funding

This research has not received not any funding from external sources.

Conflicts of interest

The authors declare no conflict of interest.

Authors' contributions statement

KB: Formulation, Data collection, Formal investigation, Methodology, Writing Codes, Development of models, Validation and Testing, and writing of original draft; PA: Writing & editing original draft.

References

[1] https://databank.worldbank.org/reports.aspx?source=2&series=SP.POP.TOTL&country=WLD#.

[2] http://www.fao.org/india/fao-in-india/india-at-a-glance/en.

[3] Zhang, K.; Wang, Q. Intelligent Systems and Decision Making for Risk Analysis and Crisis Response, Communications in Cybernetics, Systems Science and Engineering & Proceedings. Crop yield risk Assess., CRC Press; 2013, p. 709–16.

[4] Basso B, Cammarano D, Carfagna E. Review of crop yield forecasting methods and early warning systems. Proc. first Meet. Sci. Advis. Comm. Glob. Strateg. to Improv. Agric. Rural Stat. FAO Headquarters, Rome, Italy, vol. 241, 2013.

[5] Chipanshi A, Zhang Y, Kouadio L, Newlands N, Davidson A, Hill H, et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric For Meteorol 2015;206:137–50. https://doi.org/10.1016/j.agrformet.2015.03.007.

[6] López-Lozano R, Duveiller G, Seguini L, Meroni M, García-Condado S, Hooker J, et al. Towards regional grain yield forecasting with 1km-resolution EO biophysical products: Strengths and limitations at pan-European level. Agric For Meteorol 2015;206:12–32. https://doi.org/10.1016/j.agrformet.2015.02.021.

[7] Lobell DB, Thau D, Seifert C, Engle E, Little B. A scalable satellite-based crop yield mapper. Remote Sens Environ 2015;164:324–33. https://doi.org/10.1016/j.rse.2015.04.021.

[8] Zhao Y, Potgieter AB, Zhang M, Wu B, Hammer GL. Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. Remote Sens 2020;12:1024. https://doi.org/10.3390/rs12061024.

[9] Newlands NK, Zamar DS, Kouadio LA, Zhang Y, Chipanshi A, Potgieter A, et al. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. Front Environ Sci 2014;2. https://doi.org/10.3389/fenvs.2014.00017.

[10] Khaki S, Wang L. Crop Yield Prediction Using Deep Neural Networks. Front Plant Sci 2019;10. https://doi.org/10.3389/fpls.2019.00621.

[11] Kim N, Ha K-J, Park N-W, Cho J, Hong S, Lee Y-W. A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. ISPRS Int J Geo-Information 2019;8:240. https://doi.org/10.3390/ijgi8050240.

[12] Wang AX, Tran C, Desai N, Lobell D, Ermon S. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. Proc. 1st ACM SIGCAS Conf. Comput. Sustain. Soc., New York, NY, USA: ACM; 2018, p. 1–5. https://doi.org/10.1145/3209811.3212707.

[13] P. S. MG, R. B. Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms. Appl Artif Intell 2019;33:621–42. https://doi.org/10.1080/08839514.2019.1592343.

[14] Paudel D, Boogaard H, de Wit A, Janssen S, Osinga S, Pylianidis C, et al. Machine learning for large-scale crop yield forecasting. Agric Syst 2021;187:103016. https://doi.org/10.1016/j.agsy.2020.103016.

[15] Shahi S, Mousavi SF, Hosseini K. Simulation of Pan Evaporation Rate by ANN Artificial

Intelligence Model in Damghan Region. J Soft Comput Civ Eng 2021;5:75–87. https://doi.org/10.22115/scce.2021.286933.1321.

[16] Khosla E, Dharavath R, Priya R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. Environ Dev Sustain 2020;22:5687–708. https://doi.org/10.1007/s10668-019-00445-x.

[17] Kim N, Na S-I, Park C-W, Huh M, Oh J, Ha K-J, et al. An Artificial Intelligence Approach to Prediction of Corn Yields under Extreme Weather Conditions Using Satellite and Meteorological Data. Appl Sci 2020;10:3785. https://doi.org/10.3390/app10113785.

[18] Gandhi FR, Patel JN. Combined Standardized Precipitation Index and ANFIS Approach for Predicting Rainfall in the Tropical Savanna Region. J Soft Comput Civ Eng 2022;6:63–77. https://doi.org/10.22115/scce.2022.333365.1412.

[19] Cedric LS, Adoni WYH, Aworka R, Zoueu JT, Mutombo FK, Krichen M, et al. Crops yield prediction based on machine learning models: Case of West African countries. Smart Agric Technol 2022;2:100049. https://doi.org/10.1016/j.atech.2022.100049.

[20] Kuradusenge M, Hitimana E, Hanyurwimfura D, Rukundo P, Mtonga K, Mukasine A, et al. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. Agriculture 2023;13:225. https://doi.org/10.3390/agriculture13010225.

[21] Ahmed S. A Software Framework for Predicting the Maize Yield Using Modified Multi-Layer Perceptron. Sustainability 2023;15:3017. https://doi.org/10.3390/su15043017.

[22] Jhajharia K, Mathur P, Jain S, Nijhawan S. Crop Yield Prediction using Machine Learning and Deep Learning Techniques. Procedia Comput Sci 2023;218:406–17. https://doi.org/10.1016/j.procs.2023.01.023.

[23] Durrant A, Markovic M, Matthews D, May D, Leontidis G, Enright J. How might technology rise to the challenge of data sharing in agri-food? Glob Food Sec 2021;28:100493. https://doi.org/10.1016/j.gfs.2021.100493.

[24] Quinlan JR. Learning with continuous classes. 5th Aust. Jt. Conf. Artif. Intell., vol. 92, World Scientific; 1992, p. 343–8.

[25] Nourani V, Tajbakhsh AD, Molajou A. Data mining based on wavelet and decision tree for rainfall-runoff simulation. Hydrol Res 2019;50:75–84. https://doi.org/10.2166/nh.2018.049.